2017

The Use of Informativity in the Development of Robust Viromics-based Examinations

Siobhan C. Watkins
Loyola University Chicago

Catherine Putonti
Loyola University Chicago, cputonti@luc.edu

Recommended Citation

Watkins SC, Putonti C. (2017) The use of informativity in the development of robust viromics-based examinations. PeerJ 5:e3281 https://doi.org/10.7717/peerj.3281

The use of informativity in the development of robust viromics-based examinations

Siobhan C. Watkins[1,2] and Catherine Putonti[2,3,4,5]

[1] Biology Department, New Mexico Institute of Mining and Technology, Socorro, NM, United States of America
[2] Department of Biology, Loyola University of Chicago, Chicago, IL, United States of America
[3] Department of Computer Science, Loyola University of Chicago, Chicago, IL, United States of America
[4] Bioinformatics Program, Loyola University of Chicago, Chicago, IL, United States of America
[5] Department of Microbiology and Immunology, Loyola University of Chicago, Maywood, IL, United States of America

ABSTRACT

Metagenomics-based studies have provided insight into many of the complex microbial communities responsible for maintaining life on this planet. Sequencing efforts often uncover novel genetic content; this is most evident for phage communities, in which upwards of 90% of all sequences exhibit no similarity to any sequence in current data repositories. For the small fraction that can be identified, the top BLAST hit is generally posited as being representative of a viral taxon present in the sample of origin. Homology-based classification, however, can be misleading as sequence repositories capture but a small fraction of phage diversity. Furthermore, lateral gene transfer is pervasive within phage communities. As such, the presence of a particular gene may not be indicative of the presence of a particular viral species. Rather, it is just that: an indication of the presence of a specific gene. To circumvent this limitation, we have developed a new method for the analysis of viral metagenomic datasets. BLAST hits are weighted, integrating the sequence identity and length of alignments as well as a taxonomic signal, such that each gene is evaluated with respect to its information content. Through this quantifiable metric, predictions of viral community structure can be made with confidence. As a proof-of-concept, the approach presented here was implemented and applied to seven freshwater viral metagenomes. While providing a robust method for evaluating viral metagenomic data, the tool is versatile and can easily be customized to investigations of any environment or biome.

Subjects Bioinformatics, Computational Biology, Microbiology, Virology
Keywords Virome, Metagenomics, Bacteriophage, Viral community

Submitted 12 September 2016
Accepted 7 April 2017
Published 2 May 2017

Corresponding author
Catherine Putonti, cputonti@luc.edu

Academic editor
Thomas Rattei

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.3281**

BACKGROUND

Bacterial viruses (bacteriophages) play a crucial role in shaping microbial populations and processes on a global scale. They shape community structure via mediation of mortality and drive diversity as agents of genetic mobility (*Wilhelm & Suttle, 1999*; *Canchaya et al., 2003*; *Berdjeb et al., 2011*; *Clokie et al., 2011*; *Winget et al., 2011*; *Willner et al., 2012*; *Brum et al., 2016*; *Manrique et al., 2016*), and their impact has been described at higher trophic levels (*Rohwer & Thurber, 2009*; *Jover et al., 2014*). Despite being the most ubiquitous and abundant biological entity on the planet, only a comparatively small fraction of phage

genomes has been sequenced (*Klumpp, Fouts & Sozhamannan, 2012*). Nevertheless, from this small and imprecise representation of phage diversity we have uncovered a great deal about their genomes: they span a remarkable degree of genetic diversity and often have highly mosaic genome architectures (*Hatfull, 2008*; *Hatfull, 2015*). The majority of phage genes, however, are unfamiliar to us, their function unknown (*Hatfull, 2008*; *Sharon et al., 2011*). Nevertheless, as is true of all aspects of microbial diversity in the environment, the significance of the work performed to date does not negate how much there is left to discover.

Numerous studies of phage communities spanning a wide variety of environments, from the human gut (*Minot et al., 2013*) to terrestrial hot springs (*Gudbergsdóttir et al., 2015*), have repeatedly found that we are underestimating the genetic diversity within phage populations (*Dinsdale et al., 2008*; *Halary et al., 2010*; *Hurwitz & Sullivan, 2013*; *Paez-Espino et al., 2016*). Conserved taxonomic "gene signature" sequences (e.g., g20 (*Short & Suttle, 2005*) and g23 (*Filée, Tétart & Krisch, 2005*)) are far from comprehensive (*Adriaenssens & Cowan, 2014*); and there are likely groups in nature that do not contain a single signature gene identified within existing clades. Thus, whole genome sequencing (WGS) is widely considered to be the most representative method for exploring viral diversity in the environment. Bioinformatic approaches for analyzing viral metagenomes largely mirror those used for the study of bacterial and archaeal populations: reads or contigs are compared to known, characterized sequences within public data repositories. While comparisons can be made to, e.g., all viral genome sequences, another option is direct comparison to Prokaryotic Virus Orthologous Groups (pVOGs, formerly called Phage Orthologous Groups, POGs) (*Kristensen et al., 2010*; *Kristensen et al., 2013*; *Grazziotin, Koonin & Kristensen, 2017*), including 57 taxon-specific "signature" sequences (*Kristensen et al., 2013*). This approach has been employed frequently (e.g., *Kristensen et al., 2010*; *Waller et al., 2014*; *Jeffries et al., 2015*; *Laffy et al., 2016*) and these taxon-specific signatures include genes that are not found in genomes of other viral taxa. But the diversity of phages is severely undersampled, and therefore it is not surprising then that only a small fraction of sequences from viral metagenomic surveys exhibit any homology to extant databases or these signature sequences (*Hurwitz & Sullivan, 2013*; *Bruder et al., 2016*; *Paez-Espino et al., 2016*).

For the few viral species that can be identified, typically via BLAST searches against complete viral genomes or the aforementioned POG/pVOG sequences, the best hit is often regarded as being representative of the viral taxon containing the homologous region (particularly if the hit is to one of the taxon-specific signatures). This approach is employed by many metagenomics-based studies, analytical tools, and metrics (e.g., *Wommack et al., 2012*; *Huson & Weber, 2013*; *Roux et al., 2014*; *Aziz et al., 2015*; *Keegan, Glass & Meyer, 2016*). Homology-based classifications, however, can be misleading due to two factors. Firstly, phage genomes available in public repositories: (a) capture but a small fraction of the viral diversity on Earth, (b) represent phages with hosts amicable to growth under laboratory conditions, and (c) phage groups have very biased sampling rates (e.g., the heavily sampled Mycobacteriophage vs. the less-sampled phages of *Burkholderia*) (*Bruder et al., 2016*). Secondly, lateral gene transfer (LGT) is pervasive within phages communities.

There is an abundance of evidence of LGT between phages with similar host ranges, between phages within the same environment, and between phages and their hosts (e.g., *Mann et al., 2003*; *Brussow, Canchaya & Hardt, 2004*; *Lindell et al., 2005*; *Lima-Mendez et al., 2008*; *Thompson et al., 2011*; *Gao, Gui & Zhang, 2012*).

Here, we introduce a rigorous method for classifying viromes. Genes exhibiting homology to characterized sequences are weighted based upon their *informativity*—a new metric for describing viral community structure. This metric provides a means for distinguishing (and qualifying this distinction) between the presence/absence of a particular taxonomical group and genic content. Thus, it is possible to distinguish between genes indicative of a particular taxa and those that are frequently exchanged within viral communities. In addition to presenting the method, we have tested its robustness through the analysis of all individual genera of tailed bacteriophages (order: *Caudovirales*). As a proof-of-concept, we examined seven publicly available freshwater DNA metagenomic datasets.

MATERIALS AND METHODS

Development of the informativity metric

Establishing a taxonomic signal threshold

To ascertain the presence/absence of a specific taxon within a metagenome, we suggest a threshold to differentiate between informative and uninformative hits. The taxonomic signal threshold T is determined through a two-step process prior to evaluation of the metagenomic data. In the first step, each annotated coding region for a given taxon of interest is compared to all annotated sequences within the genome(s) of a known relative. Thus, each coding region's sequence x ($x \in X$, where X is the set of sequences for all coding regions annotated within the genome of the taxon of interest) is compared to each coding region's sequence g ($g \in G$, where G is the set of sequences for all coding regions annotated within the genome of a known relative). The use of a known relative genome(s) establishes if and how conserved the coding region is between known, related strains/species. Where sequence homology is detected, the sequence identity and query coverage of the match is recorded: S_1 and Q_1, respectively.

In the second step, each coding region's sequence is compared again, this time to the sequences for all annotated coding regions for the group assayed by the metagenome (e.g., all phages, viruses, bacteria, archaea, etc.), however, those belonging to the taxonomic group containing the taxon of interest and the known relative considered in step one are omitted. Many hits may be recorded for a particular gene x. Thus the best hit, the highest scoring hit both with respect to the sequence identity and the query coverage of the match, is selected; S_2 and Q_2 denote this best match's sequence identity and query coverage, respectively. A taxonomic signal threshold T is defined as $T = \{S_1 - S_2, Q_1 - Q_2\}$ where the subscripts *1* and *2* represent the sequence identity and query coverage of the match detected from steps one and two, respectively. Figure 1 illustrates the two-step process and the T values produced.

It is important to note that the taxonomic group used for comparison is user defined. For instance, in order to ascertain if a gene can be used to distinguish between the

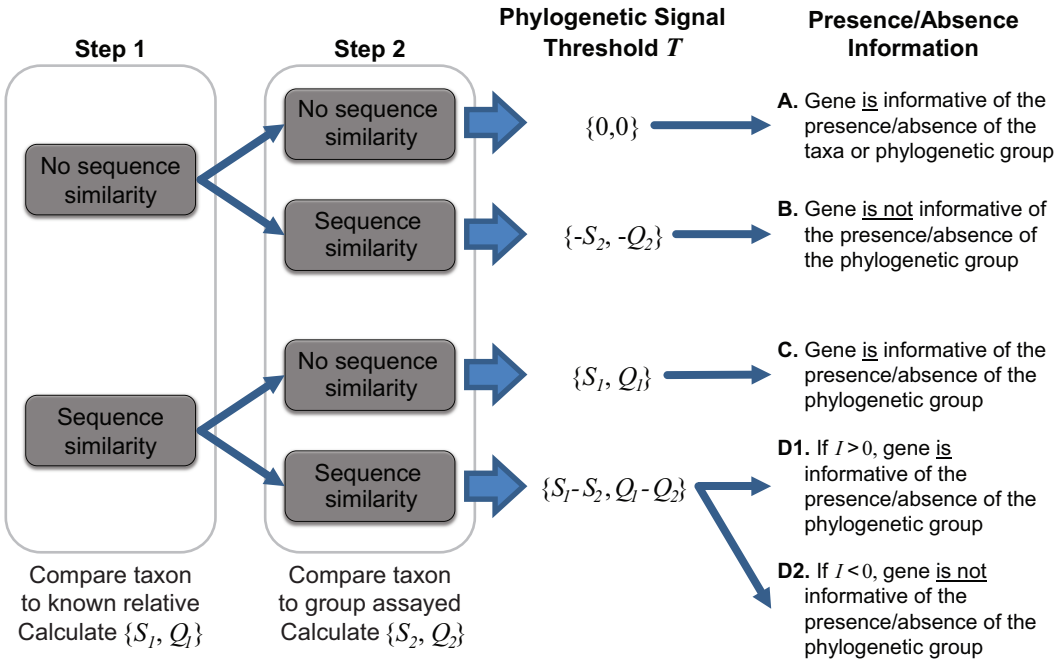

Figure 1 Two-step process for determining the taxonomic signal threshold T and the information which can be gained regarding the presence/absence of a taxon's phylogenetic group. S_1 and S_2 represent the sequence identity of homologies identified in step 1 and 2, respectively. Likewise, Q_1 and Q_2 refer to the query coverage of the match detected in step 1 and 2, respectively.

presence/absence of a particular species, one may consider the taxonomic group to be inclusive only of strains of the species. Therefore, in this case, the most distant relative belonging to the taxonomic group in step one would be the closest related species. If a more distant relative, say the most distantly related species of the same genus, were to be investigated, then the taxonomic signal threshold T would serve as a means to distinguish between the presence/absence of a subset of the species (inclusive of the taxon of interest) within the genus. This flexibility enables the researcher to define and control the granularity of his/her analyses. If a particular taxa of interest lacks available genomes capturing the phylogenetic diversity of the species (or genus or subfamily, etc.), a more distant relative can be selected. In addition to the intended purpose of establishing the taxonomic signal threshold, the two-step process can provide insight into putative horizontally acquired elements and gene loss events, e.g., instances in which the gene did not include a homolog in the most distant relative but did exhibit sequence similarity to a gene within the genome of another taxonomic group.

Using informativity to ascertain confidence in taxonomical calls

As indicated in Fig. 1, when T is greater than zero (outcomes C and D1), the presence of a specific gene can provide insight. Operational Taxonomic Unit (OTU) calls are informed by this threshold to decipher BLAST analyses of metagenomic datasets as some hits may be to genes which are conserved and thus poor indicators if a species/taxa is present or absent. For a given hit within a metagenomic dataset, the sequence identity and query coverage, S_H

and Q_H respectively, is assessed relative to the taxonomic signal threshold T for the gene producing the match. Genes in which T <0 have already been classified as uninformative (Fig. 1). Hits which fall below the gene's threshold, $\{S_H, Q_H\} < T$, are also classified as uninformative, while hits which are above the threshold are considered informative. The informativity I of each hit is quantified based upon deviation from this threshold T such that $I = \{S_H, Q_H\}$-T. I can range from 0 (equivalent to the threshold T) to 100 ($T = \{0,0\}$, $S_H = Q_H = 100\%$). Thus, genes with a high value of I are strong indicators of the presence of the gene from the taxon of interest (or a closely related strain/species) within a metagenomic dataset.

Taking into consideration the number of informative genes detected within a metagenomic sample and their individual I values, one can then quantify with confidence the likelihood of the presence/absence of the taxon of interest. For example, consider the case in which a novel species, *n*, within a genus is represented within a metagenome. It shares homology with other genomes for the genus. For the sake of simplicity assume there are two other genomes for the genus: *a* and *b*. The novel species *n*'s genome contains a subset of genes that are more similar to informative genes in *a*'s genomes and some genes that are more similar to informative genes in *b*'s genome. One can use the informativity values calculated for the genes of *n* to provide a confidence value in calling the contig a representative of *a* and/or *b*. Furthermore, rather than simply assign the contig as a representative of *a* or *b* or simply a member of a particular genus based upon a single signature gene, the informativity metric can provide insight into the evolutionary history of this novel species and the taxa.

Implementation

The method for assessing the informativity of viromic hits was implemented using a series of BLAST databases and BLAST searches. A collection of all coding regions (nucleotide sequences) for the taxon of interest (X) and all genes (amino acid sequences) annotated within the genome of the selected relative (G) are supplied by the user. A local BLAST database is created for *G*, and the genes belonging to X are queried against the local database via blastx. The sequence identity and query coverage of the match detected for the best hit for each gene is then parsed from the BLAST results quantifying each gene's S_1 and Q_1 values. Next, a BLAST database is created for the annotated coding regions (amino acid sequences) provided for step 2 of this method (set Z), again supplied by the user. Each of the genes for the taxon of interest X is queried against this second local database via blastx; the results are again parsed for each gene's S_2 and Q_2 values so that the taxonomic signal threshold T can be calculated.

A metagenomic dataset can next be evaluated, comparing each read or contig against a collection of annotated gene sequences. To accommodate the variation between characterized sequences in databases and environmental samples, contigs are translated—generating all six open reading frames—and a protein database representative of the metagenomic dataset is produced. Each BLAST hit is next assessed with respect to its scores $\{S_H, Q_H\}$ relative to that of the gene's threshold T. For each gene in the genome of interest X, the values for S_1, Q_1, S_2, Q_2, S_H, and Q_H are written to file. The user can

then evaluate the likelihood of a particular taxon's or taxonomic group's presence within the metagenomic sample based upon the I values for informative genes. Note that for the analyses presented here we have weighted S and Q values equally in the calculation of T; the two values are, however, reported separately such that users can select their own weighting of the contributions of sequence identity and query coverage.

The described process has been automated via functionality developed in C++ (available for both Windows and Unix OS). Users must supply or specify the FASTA format files for the taxon of interest (X), the known relative (G), and the group assayed (less the taxonomic group of interest) (Z). If metagenomic comparisons are to be conducted, as this is optional in the current implementation, the user must also supply the metagenomic dataset. The code has been designed for both ease of use, speed, and flexibility, such that analyses can be tailored to the environmental niche and/or hypothesis under investigation. Most importantly, this is a light-weight solution which can be integrated into the standard method of viral metagenomic analyses. Source code, documentation, and sample data are publicly available at https://github.com/putonti/informativity.

Datasets examined

Viral gene and genome datasets

Sequence data were retrieved from NCBI GenBank (*NCBI Resource Coordinators, 2017*) (collected August 2016). Datasets for 70 taxonomical groups within *Caudovirales* were retrieved (Table S1); searches were conducted in NCBI for protein sequences through an advanced search query: PHG[Division] AND txidXXXXX[Organism] (where the X's refer to the NCBI Taxonomy Browser's Taxonomy ID number). Note, this only collects phages that have been annotated to the taxon (i.e., their genome has been annotated with the Taxonomy ID). From these queries, 70 sets of genome sequences were retrieved. Sixty-four individual genera were selected. The other six sets consist of sequences for species belonging to the same subfamily. *Caudovirales* taxa were selected as they are the largest and best characterized phage genomes currently available (*Salmond & Fineran, 2015*). In addition, phages classified within other orders were retrieved with the following query: (PHG[Division] NOT txid28883[Organism]); Taxonomy ID 28883 is the unique identifier for *Caudovirales*. The results of this query include all phages belonging to other orders (1,003 phage strains in total). For each *Caudovirales* taxonomical group, the type species' genome was retrieved, again from NCBI. The type species was determined by referring to the 2015 release by the International Committee on Taxonomy of Viruses (ICTV) (http://www.ictvonline.org). The type species for each *Caudovirales* taxonomical group is listed in Table S1 .

In our proof-of-concept analyses of the *Pbunavirus Pseudomonas phage PB1*, we verified the taxonomic classification of Pbunaviruses. Genomes exhibiting significant homology (>50% of coding regions) to PB1 that were not assigned to the *Pbunavirus* Taxonomy ID were further investigated. The complete sequence of the genome in question was aligned via the blastn algorithm through the NCBI BLAST site. Alignments with a query coverage and percent identity greater than 50% were identified and the literature was referenced to correctly assign the taxonomic classification. Additional *Pbunavirus* strains were identified

from the "unclassified *Myoviridae*" following this above method. These genomes were thus reannotated for our subsequent analysis of viral metagenomic datasets as *Pbunavirus*. (See Table S2 for a list of the genomes classified here as Pbunaviruses.) Pbunaviruses were selected for this proof-of-concept work given our prior isolation and identification of *Pseudomonas phage PB1* in the freshwaters of Lake Michigan (*Malki et al., 2015*).

Viral metagenomic analyses

SRA records were collected from the NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra). Table S3 lists all of the datasets included in the proof-of-concept study. Each SRA record (line listed in the Table S3) was considered as an individual sample with two exceptions. Two samples are aggregates of more than one SRA record, both belonging to Virome IV, as they were combined in the downloadable file from the SRA database. Our dataset includes 56 individual samples. These samples were chosen as they target DNA viruses in similar environments (freshwater). Furthermore, they are rather well documented in the literature. Each individual sample was next assembled using Velvet (*Zerbino & Birney, 2008*) with a hash size of 31; default values were used for all other parameters. Each sample was thus uniformly prepared for analysis.

The amino acid and nucleotide sequences for *Pseudomonas phage PB1* (type strain for the *Pbunavirus* genus; Accession Number: NC_011810) and *Burkholderia phage BcepF1* (Accession Number: NC_009015) were downloaded from NCBI for comparison with the virome datasets. All phage nucleotide sequences (omitting those belonging to the *Pbunavirus*) were also retrieved via the advanced search query: PHG[Division] NOT txid1198980[Organism] (where the Taxonomy ID listed is that for *Pbunavirus*). In total over 500000 individual records were retrieved, including partial and complete sequences. The informativity values are visualized in later figures as heatmaps that were produced in Excel.

RESULTS AND DISCUSSION

Identifying informative genes

The new metric described here, Informativity or *I*, provides a quantifiable means of identifying if a particular taxonomical group is present/absent within a sequenced community. Developed specifically for the detection of viral sequences in complex community metagenomic data sets, *I* captures the likelihood of a sequence belonging to taxa. Described in greater detail within the Methods, Fig. 2 provides an overview of how informative genes are identified. Users must supply the query sequence(s) (likely a contig or set of contigs from a sequenced community), at least two representative sequences for a taxon of interest, and lastly a set of sequences representative of 'non-relatives' (sequences belonging to other taxa of, e.g., viruses). The taxon of interest can be, e.g., a species, a genus, or a subfamily.

Informative genes for *Caudovirales* taxa

All protein coding sequences were collected for species belonging to 70 tailed-virus (*Caudovirales*) taxa identified by NCBI Taxonomy (see Methods). Using the ICTV type

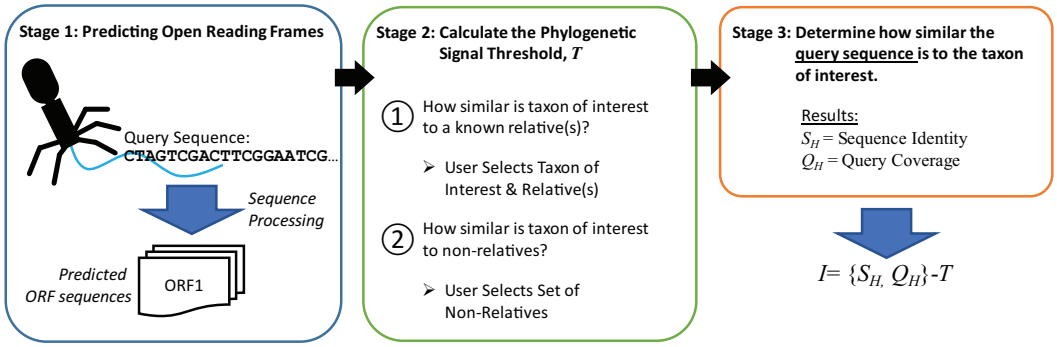

Figure 2 Process for calculating informativity. In Stage 1, users supply their assembled contigs which are processed, predicting ORFs. Users must supply at minimum two sequences for the taxon of interest (preferably spanning the diversity of sequences within the taxon) and sequences of 'non-relatives' for the calculation of the phylogenetic signal threshold T in Stage 2. Each gene's informativity is calculated in Stage 3.

species as a representative of the taxa, each gene sequence (x) of the type species' genome (X) was compared to all other gene sequences for species of the same taxa. For each gene, the sequence identity S_1 and query coverage of the match Q_1 for the most dissimilar homologous gene sequence within the taxa is calculated. This captures the sequence variation for the gene within the species of the taxon. Thus, the S_1 and Q_1 scores for one gene x_i may be from homology detected in one species of the taxa, while the scores for another gene x_j may be to a homolog in another species' genome. If the gene is unique to the type species' genome, then $S_1 = Q_1 = 0$. The sequence identity, S_2, and query coverage, Q_2, scores were next calculated for each gene in the type species' genome; each gene was compared to: (1) genes belonging to species classified within other genera within the order *Caudovirales* and (2) genes belonging to species of other taxonomic orders. In contrast to the S_1 and Q_1 scores, the S_2 and Q_2 scores are for the best hit or the most similar homolog found. Using these two values, the taxonomic signal threshold T can be calculated (see 'Methods'). This threshold value signifies how reliable the particular gene is as an indicator of the presence/absence of the species. Genes which are found in multiple species and taxa would thus have a low threshold value T and perform poorly as an indicator of the taxon.

Figure 3 illustrates the thresholds for *Myoviridae* and *Podoviridae* type species; *Siphoviridae* is included in Fig. S1. (Type species names and accession numbers as well as scores are listed in Table S1). In these maps, each gene's taxonomic signal threshold is shown; dark gray boxes indicate uninformative genes; these uninformative genes either exhibit greater homology to species belonging to other phage taxa or lack homology to other representative genomes of the taxon of interest (i.e., are present only within the type species' genome). Also listed for each taxon is the number of genome sequences included in the comparisons. Those taxonomical groups with more phylogenetic diversity represented within available genome sequences tend to have less informative genes. This is quite prominent when evaluating the 10 *Podoviridae* taxon: the well sampled subfamily of *Autographivirinae* species have significantly less informative genes than the undersampled *Podoviridae* genera of, e.g., F116virus and Bpp1virus. It is important to note, however, that

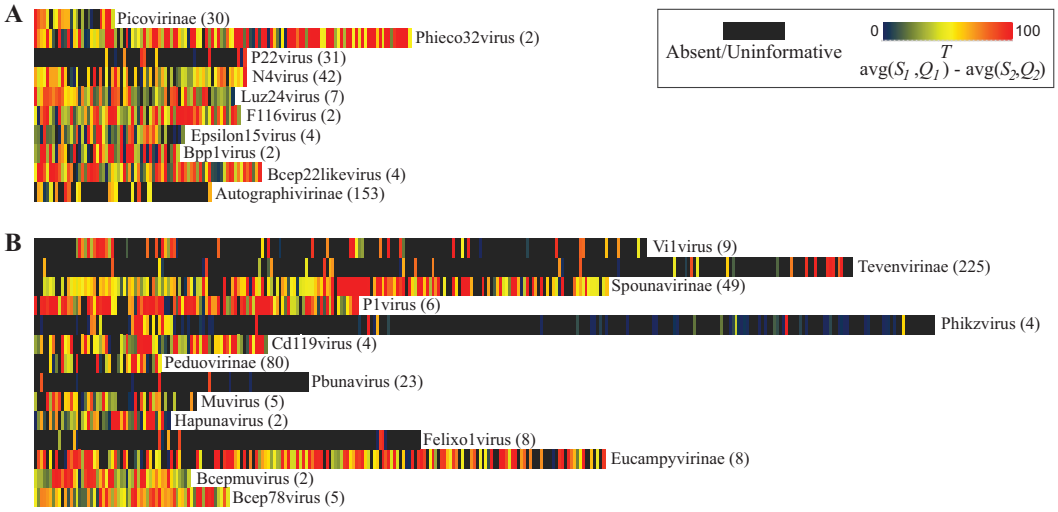

Figure 3 Taxonomic signal threshold value *T* for each gene within phage type species of taxonomic groups belonging to the family *Podoviridae* and *Myoviridae*. For each taxonomical group belonging to the viral family (**A**) *Podoviridae* and (**B**) *Myoviridae*, the number of genome sequences examined (including the type strain) is indicated in parentheses.

taking into consideration numerous genome sequences does not necessarily mean that the phylogenetic diversity of the taxon was examined. In contrast to classifying unknown sequences by a single marker, the informativity metric provides a multiple gene marker strategy. Thus, taxonomical 'calls' for a sequence can be made with greater confidence by reporting the aggregate of informative markers found, not just the presence/absence of a single gene.

Targeting specific phages in environmental samples

The *Pseudomonas phage PB1* was selected for examination. Each gene annotated for the PB1 genome (Accession Number: NC_011810) (*Ceyssens et al., 2009*) was compared first to the set of genes for the most distant relative of PB1 within its genus *Pbunavirus* (previously called *Pbunalikevirus*), *Burkholderia phage BcepF1* (Accession Number: NC_009015). For each gene the S_1 and Q_1 values were computed. Next, all 93 annotated PB1 genes were compared to all genes from phage species—other than those classified as Pbunaviruses (see 'Methods'). Homologous sequences were identified, the S_2 and Q_2 values. The similarity observed (the S_2 and Q_2 values) for each of the PB1 genes is shown in the heatmap of Fig. 4. Several PB1 gene sequences (as indicated by the color scale) exhibited sequence homology to genes within phage genomes of other taxa. Dark gray blocks in the heatmap signify that no homologs were detected. The upper chart in Fig. 4 details the percent sequence identity (bars) and percent query coverage (circles) values observed for the best hits to GenBank records. PB1 genes with homologies to other phage taxa include conserved genes (e.g., gp47 = tail fiber component and gp50 = DNA ligase), amongst other conserved "hypothetical proteins".

The methodology developed here was then applied to seven freshwater DNA viromes (Table 1); a list of the SRA datasets from each study is provided in Table S3. Each of the 56

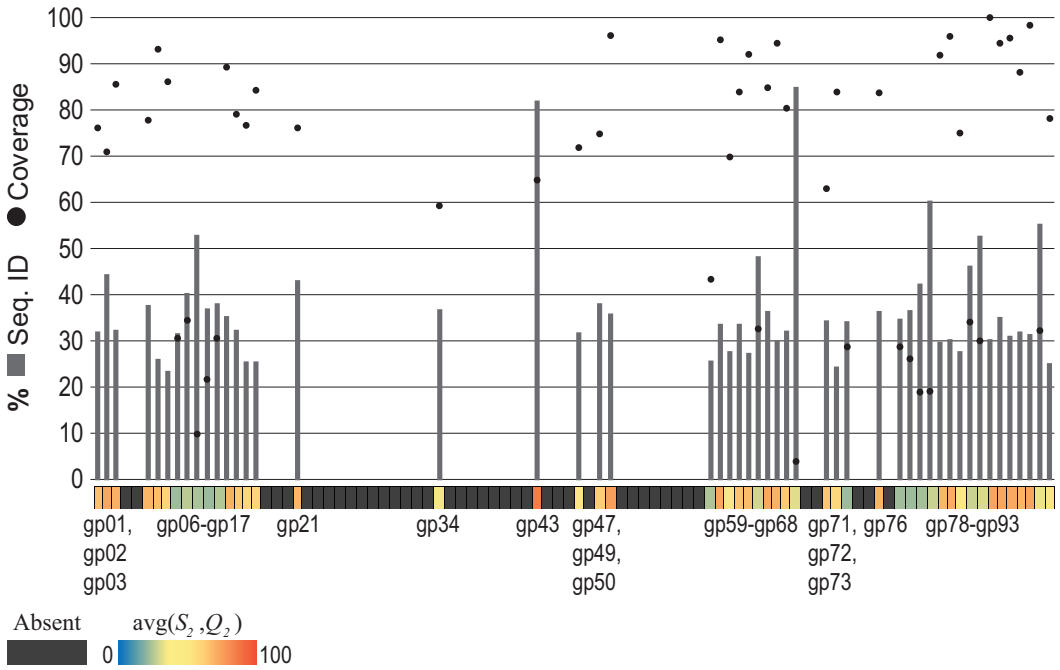

Figure 4 Observed similarity of each *Pseudomonas phage PB1* gene to phage genes within other (non-*Pbunavirus*) taxa. The percent sequence identity (bars) and percent query coverage (circles) values for the best hit for each gene is shown as is the average of these two percentages within the heatmap along the *x*-axis. Genes which do not show homology to non-Pbunaviruses are indicated as dark gray boxes within the heatmap.

Table 1 Freshwater DNA viral metagenomic studies retrieved from NCBI's SRA database.

Virome	Environmental niche	Number of samples	Sequencing technology	Mbp total	Reference
I	Lake Michigan nearshore	40	Illumina	6,909	*Watkins et al. (2015)*, *Sible et al. (2015)*
II	Lake Bourget	2	454	698	*Roux et al. (2012)*
III	Kent SeaTech tilapia pond	3	454	47	*Dinsdale et al. (2008)*
IV	Lake Limnopolar	2	454	18	*López-Bueno et al. (2009)*
V	Reclaimed water samples	6	454	364	*Rosario et al. (2009)*
VI	Lake Ontario	3	454	223	n/a
VII	Feitsui Reservoir	5	454	86	*Tseng et al. (2013)*

samples examined was first assembled (see 'Methods' for details). The PB1 coding regions were then compared to the 56 collections of contigs. The heatmap shown in Figure 5A graphically represents these results; each row represents a single sample (Methods). Again, each gene's best hit within each virome's sample was qualified (colored) with respect to its conservation amongst the Pbunaviruses, the gene's S_1 and Q_1 value. Nevertheless, not all genes provide an equal signal as to the presence or absence of PB1 within the sample: some serve as better markers. As shown in Fig. 4, there are several "non-*Pbunavirus*" species which contain homologs to PB1 genes. Thus, the informativity I of each BLAST hit within the seven viromes was calculated. In doing so, individual genes that provide a strong signal

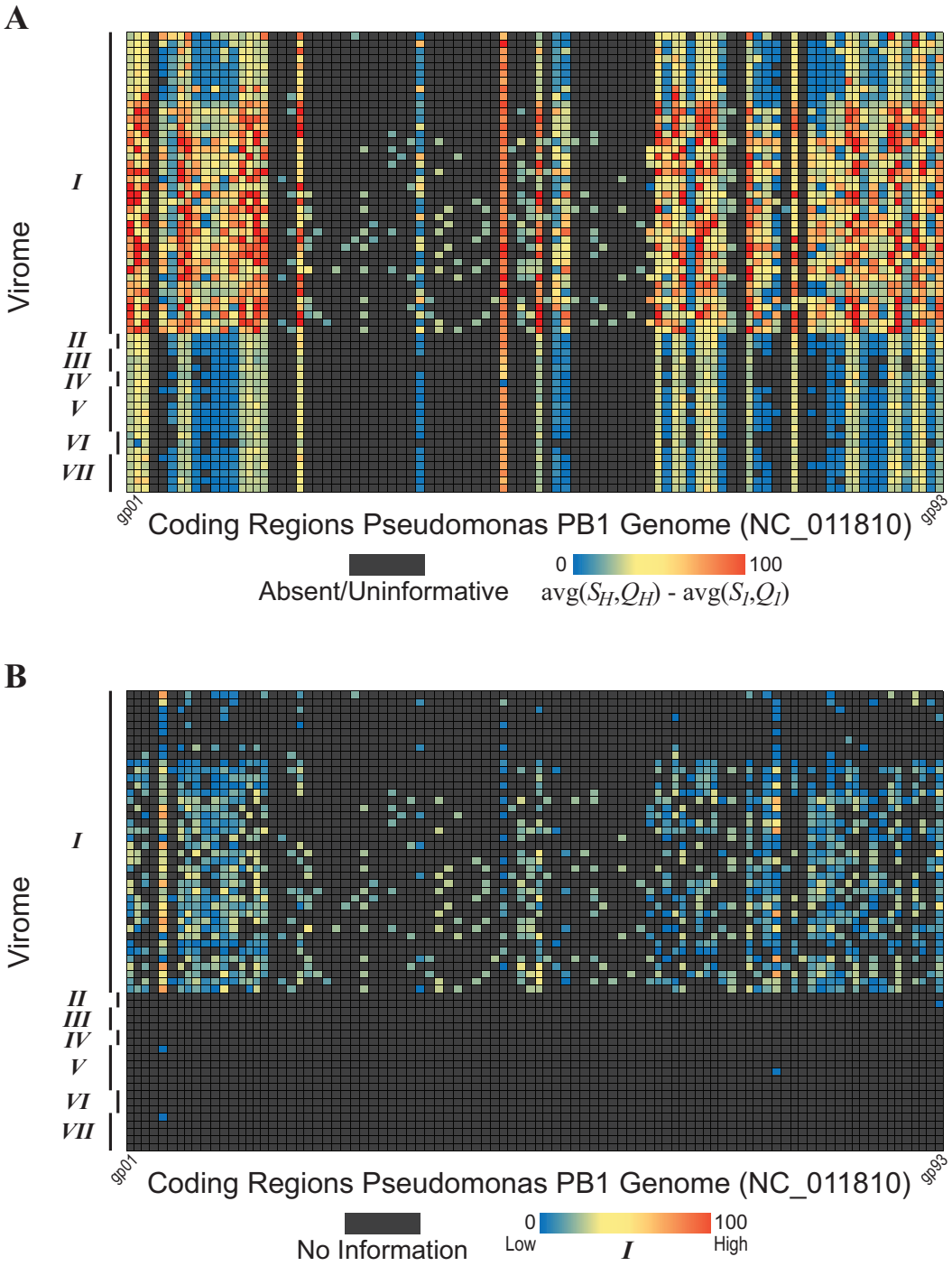

Figure 5 **Evidence of *Pseudomonas phage PB1* genes within seven freshwater DNA viromes** The seven viromes correspond to those listed in Table 1. (A) Hits (S_H and Q_H) to PB1 genes within viromes. As shown by the color scale, some hits to PB1 genes are better (in terms of sequence identity and query coverage) than homologies observed with the distant *Pbunavirus Burkholderia phage BcepF1*. (B) Hits are qualified relative to the taxonomic signal threshold T calculated for PB1 genes.

for the *Pseudomonas phage PB1* can readily be identified. Figure 5B represents the results of this computation, in which each hit to a PB1 gene is now assessed in light of the taxonomic signal threshold T.

In an effort to assess the strength of the metric presented here, we evaluated the raw BLAST results of the datasets and a BLAST score-based analysis. The BLAST results of Viromes II, IV, V, and VII are publicly available through the web service MetaVir (*Roux et al., 2014*). Nine of the samples from Virome I are also available through MetaVir. It is important to note that in contrast to the uniform method in which the viral metagenomes were preprocessed here (see 'Methods'), the sequences submitted to MetaVir, or comparable online resources, may be assembled or raw sequences. Furthermore, MetaVir conducts BLAST comparisons against the RefSeq viral database (*O'Leary et al., 2016*), whereas here we have included all partial and complete phage sequences from GenBank. Nevertheless, hits to the *Pbunavirus* (Table S2) genomes were identified in all five MetaVir datasets; the Lake Michigan and Lake Bourget samples (nine samples from Virome I and both samples from Virome II, respectively) produced the most hits in MetaVir to the *Pbunavirus* genomes (hundreds to thousands), many which were the best hits identified. Hits from MetaVir metagenomic samples, including Viromes I, II, IV, V, VII and additional sampling sites not included in our proof-of-concept work, to the *Pseudomonas phage PB1* genome are shown in Fig. S2.

Virome I, the Lake Michigan viral metagenomes generated by our group (*Watkins et al., 2015*; *Sible et al., 2015*), includes many informative genes (Fig. 5B) indicative of the presence of a *Pbunavirus* similar to PB1. Thus, with confidence, one can predict its presence within this sample. Viromes II, V, and VII contain far fewer hits to informative genes (one, two, and one PB1 genes respectively). Furthermore, their informativity scores are low, $\{S_H, Q_H\} \approx T$. This would suggest that PB1 (or a close relative) is not present within the sample: rather a homolog of the gene is present, within an uncharacterized species. As viral sequence databases expand through the isolation and characterization of additional viruses, the threshold T is likely to change thus providing greater confidence in the evaluation of BLAST hits for OTU calling.

CONCLUSIONS

The method presented here, for extrapolating the presence/absence of microbial taxa, is robust and versatile. By scrutinizing a set of informative genes, the effects of lateral gene transfer and incomplete, sparse databases are reduced. Furthermore, as new genome sequences are released, the informativity metric can be easily updated. Specifically, the proof-of-concept investigation of seven freshwater virome datasets can be applied to identify novel strains and species of phages with confidence and thus easily mine large datasets for specific taxa of interest. Many of the cellular constituents of the human microbiome are undergoing examination, and exploration of human viromes is certainly the next frontier (*Abeles & Pride, 2014*; *Ogilvie & Jones, 2015*; *Handley, 2016*; *Manrique et al., 2016*; *Zou et al., 2016*). These studies have already discovered novel phage species (*Dutilh et al., 2014*; *Malki et al., 2016*) and will undoubtedly continue to increase our

understanding of phage diversity. Nevertheless, improved bioinformatic tools for mining sequences representative of complex viral communities, coupled with further physical isolation and characterization of viral species have the potential to greatly expand our knowledge of the viral diversity on Earth.

ACKNOWLEDGEMENTS

The authors would like to thank Ms. Katherine Bruder, Alexandria Cooper, Kema Malki, and Emily Sible for their contributions to general research investigating Pbunaviruses. Thanks also to Mr. Thomas Hatzopoulos for his contributions during early code development.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the NSF (award #1149387) to CP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NSF: 1149387.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Siobhan C. Watkins conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Catherine Putonti conceived and designed the experiments, performed the experiments, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper, and was responsible for code development.

Data Availability

The following information was supplied regarding data availability:
GitHub: https://github.com/putonti/informativity.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3281#supplemental-information.

REFERENCES

Abeles SR, Pride DT. 2014. Molecular bases and role of viruses in the human microbiome. *Journal of Molecular Biology* **426**:3892–3906 DOI 10.1016/j.jmb.2014.07.002.

Adriaenssens EM, Cowan DA. 2014. Using signature genes as tools to assess environmental viral ecology and diversity. *Applied and Environmental Microbiology* **80**:4470–4480 DOI 10.1128/AEM.00878-14.

Aziz RA, Dwivedi B, Akhter S, Breitbart M, Edwards RA. 2015. Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. *Frontiers in Microbiology* **6**:Article 381 DOI 10.3389/fmicb.2015.00381.

Berdjeb L, Pollet T, Domaizon I, Jacquet S. 2011. Effect of grazers and viruses on bacterial community structure and production in two contrasting trophic lakes. *BMC Microbiology* **11**:88 DOI 10.1186/1471-2180-11-88.

Bruder K, Malki K, Cooper A, Sible E, Shapiro JW, Watkins SC, Putonti C. 2016. Freshwater metaviromics and bacteriophages: a current assessment of the state of the art in relation to bioinformatic challenges. *Evolutionary Bioinformatics* **12**:25–33 DOI 10.4137/EBO.S38549.

Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. 2016. Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME Journal* **10**:437–449 DOI 10.1038/ismej.2015.125.

Brussow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68**:560–602 DOI 10.1128/MMBR.68.3.560-602.2004.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. 2003. Phage as agents of lateral gene transfer. *Current Opinion in Microbiology* **6**:417–424 DOI 10.1016/S1369-5274(03)00086-9.

Ceyssens P-J, Miroshnikov K, Mattheus W, Krylov V, Robben J, Noben J-P, Vanderschraeghe S, Sykilinda N, Kropinski AM, Volckaert G, Mesyanzhinov V, Lavigne R. 2009. Comparative analysis of the widespread and conserved PB1-like viruses infecting *Pseudomonas aeruginosa*. *Environmental Microbiology* **11**:2874–2883 DOI 10.1111/j.1462-2920.2009.02030.x.

Clokie MR, Millard AD, Letarov AV, Heaphy S. 2011. Phages in nature. *Bacteriophage* **1**:31–45 DOI 10.4161/bact.1.1.14942.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. 2008. Functional metagenomic profiling of nine biomes. *Nature* **452**:629–632 DOI 10.1038/nature06810.

Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5**:Article 4498 DOI 10.1038/ncomms5498.

Filée J, Tétart Suttle, CA, Krisch HM. 2005. Marine T7-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proceedings of the National Academy of Sciences of the United States of America* **30**:12471–12476 DOI 10.1073/pnas.0503404102.

Gao E-B, Gui J-F, Zhang Q-Y. 2012. A novel cyanophages with a cyanobacterial nonbleaching protein A gene in the genome. *Journal of Virology* **86**:236–245 DOI 10.1128/JVI.06282-11.

Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Research* **45**:D491–D498 DOI 10.1093/nar/gkw975.

Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X. 2015. Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. *Environmental Microbiology* **18**:863–874 DOI 10.1111/1462-2920.13079.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences of the United States of America* **107**:127–132 DOI 10.1073/pnas.0908978107.

Handley SA. 2016. The virome: a missing component of biological interaction networks in health and disease. *Genome Medicine* **8**:Article 32 DOI 10.1186/s13073-016-0287-y.

Hatfull GF. 2008. Bacteriophage genomics. *Current Opinion in Microbiology* **11**:447–453 DOI 10.1016/j.mib.2008.09.004.

Hatfull GF. 2015. Dark matter of the biosphere: the amazing world of bacteriophage diversity. *Journal of Virology* **89**:8107–8110 DOI 10.1128/JVI.01340-15.

Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLOS ONE* **8**:e57355 DOI 10.1371/journal.pone.0057355.

Huson DH, Weber N. 2013. Microbial community analysis using MEGAN. *Methods in Enzymology* **531**:465–485 DOI 10.1016/B978-0-12-407863-5.00021-6.

Jeffries TC, Ostrowski M, Williams RB, Xie C, Jensen RM, Grzymski JJ, Senstius SJ, Givskov M, Hoeke R, Philip GK, Neches RY, Drautz-Moses DI, Chénard C, Paulsen IT, Lauro FM. 2015. Spatially extensive microbial biogeography of the Indian Ocean provides insights into the unique community structure of a pristine coral atoll. *Scientific Reports* **5**:15383 DOI 10.1038/srep15383.

Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat Rev Microbiol* **12**:519–528 DOI 10.1038/nrmicro3289.

Keegan KP, Glass EM, Meyer F. 2016. MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods in Molecular Biology* **1399**:207–233 DOI 10.1007/978-1-4939-3369-3_13.

Klumpp J, Fouts DE, Sozhamannan S. 2012. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage.* **2**:190–199 DOI 10.4161/bact.22111.

Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. 2010. New dimensions of the virus world discovered through metagenomics. *Trends in Microbiology* **18**:11–19 DOI 10.1016/j.tim.2009.11.003.

Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. 2013. Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of Bacteriology* **195**:941–950 DOI 10.1128/JB.01801-12.

Laffy PW, Wood-Charlson EM, Turaev D, Weynberg KD, Botté ES, Van Oppen MJ, Webster NS, Rattei T. 2016. HoloVir: a workflow for investigating the diversity and function of viruses in invertebrate holobionts. *Frontiers in Microbiology* **7**:822 DOI 10.3389/fmicb.2016.00822.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution* **25**:762–777 DOI 10.1093/molbev/msn023.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**:86–89 DOI 10.1038/nature04111.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. *Science* **326**:858–861 DOI 10.1126/science.1179287.

Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, Watkins SC, Putonti C. 2015. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal* **12**:164 DOI 10.1186/s12985-015-0395-0.

Malki K, Sible E, Cooper A, Garretto A, Bruder K, Watkins SC, Putonti C. 2016. Seven bacteriophages isolated from the female urinary microbiota. *Genome Announc.* **4**:e01003–16 DOI 10.1128/genomeA.01003-16.

Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**:741 DOI 10.1038/424741a.

Manrique P, Bolduc B, Van der Oost J, De Vos WM, Young MJ. 2016. Healthy human gut phageome. *Proceedings of the National Academy of Sciences of the United States of America* **113**:10400–10405 DOI 10.1073/pnas.1601060113.

Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **110**:12450–12455 DOI 10.1073/pnas.1300833110.

NCBI Resource Coordinators. 2017. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **45**:D12–D17 DOI 10.1093/nar/gkw1071.

Ogilvie LA, Jones BV. 2015. The human gut virome: a multifaceted majority. *Frontiers in Microbiology* **6**:918 DOI 10.3389/fmicb.2015.00918.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A,

Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**:D733–D745 DOI 10.1093/nar/gkv1189.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. *Nature* **536**:425–430 DOI 10.1038/nature19094.

Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. *Nature* **459**:207–212 DOI 10.1038/nature08060.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. 2009. Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **11**:2806–2820 DOI 10.1111/j.1462-2920.2009.01964.x.

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLOS ONE* **7**:e33641 DOI 10.1371/journal.pone.0033641.

Roux S, Tournayre J, Mahul A, Debroas D, Enault F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**:76 DOI 10.1186/1471-2105-15-76.

Salmond GPC, Fineran PC. 2015. A century of the phage: past, present and future. *Nature Reviews Microbiology* **13**:777–786 DOI 10.1038/nrmicro3564.

Sharon I, Battchikova N, Aro E-M, Giglione C, Meinnel T, Glaser F, Pinter RY, Breitbart M, Rohwer F, Béjà O. 2011. Comparative metagenomics of microbial traits within oceanic viral communities. *ISME Journal* **5**:1178–1190 DOI 10.1038/ismej.2011.2.

Short CM, Suttle CA. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Applied and Environmental Microbiology* **71**:480–486 DOI 10.1128/AEM.71.1.480-486.2005.

Sible E, Cooper A, Malki K, Bruder K, Watkins SC, Fofanov Y, Putonti C. 2015. Survey of viral populations within Lake Michigan nearshore waters at four Chicago area beaches. *Data Brief* **5**:9–12 DOI 10.1016/j.dib.2015.08.001.

Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **108**:E757–E764 DOI 10.1073/pnas.1102164108.

Tseng C-H, Chiang P-W, Shiah F-K, Chen Y-L, Liou J-R, Hsu T-C, Maheswararajah S, Saeed I, Halgamuge S, Tang SL. 2013. Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME Journal* **7**:2374–2386 DOI 10.1038/ismej.2013.118.

Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV, Bork P. 2014. Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME Journal* **8**:1391–1402 DOI 10.1038/ismej.2014.30.

Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P, Ruggles-Sage B, Romer Z, Putonti C. 2015. Assessment

of a metaviromic dataset generated from nearshore Lake Michigan. *Marine and Freshwater Research* **67**:1700–1708 DOI 10.1071/MF15172.

Wilhelm SW, Suttle CA. 1999. Viruses and nutrient cycles in the Sea. *BioScience* **49**:781 DOI 10.2307/1313569.

Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, Lim YW, Rainey PB, Schmieder R, Youle M, Conrad D, Rohwer F. 2012. Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *American Journal of Respiratory Cell and Molecular Biology* **46**:127–131 DOI 10.1165/rcmb.2011-0253OC.

Winget DM, Helton RR, Williamson KE, Bench SR, Williamson SJ, Wommack KE. 2011. Repeating patterns of virioplankton production within an estuarine ecosystem. *Proceedings of the National Academy of Sciences of the United States of America* **108**:11506–11511 DOI 10.1073/pnas.1101907108.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6**:427–439 DOI 10.4056/sigs.2945050.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**:821–829 DOI 10.1101/gr.074492.107.

Zou S, Caler L, Colombini-Hatch S, Glynn S, Srinivas P. 2016. Research on the human virome: where are we and what is next. *Microbiome* **4**:32 DOI 10.1186/s40168-016-0177-y.